3-2008

On Information Pooling, Adaptability and Superefficiency in Nonparametric Function Estimation

T. Tony Cai
University of Pennsylvania

On Information Pooling, Adaptability and Superefficiency in Nonparametric Function Estimation

Abstract
The connections between information pooling and adaptability as well as superefficiency are considered. Separable rules, which figure prominently in wavelet and other orthogonal series methods, are shown to lack adaptability; they are necessarily not rate-adaptive. A sharp lower bound on the cost of adaptation for separable rules is obtained. We show that adaptability is achieved through information pooling. A tight lower bound on the amount of information pooling required for achieving rate-optimal adaptation is given. Furthermore, in a sharp contrast to the separable rules, it is shown that adaptive non-separable estimators can be superefficient at every point in the parameter spaces. The results demonstrate that information pooling is the key to increasing estimation precision as well as achieving adaptability and even superefficiency.

Keywords
adaptability, Bayes rules, information pooling, minimax, minimum risk inequalities, nonparametric regression, orthogonal series, separable rules, superefficiency, wavelets, white noise

Disciplines
Applied Mathematics | Statistics and Probability

On Information Pooling, Adaptability And Superefficiency in Nonparametric Function Estimation

T. Tony Cai *

Abstract

The connections between information pooling and adaptability as well as superefficiency are considered. Separable rules, which figure prominently in wavelet and other orthogonal series methods, are shown to lack adaptability; they are necessarily not rate-adaptive. A sharp lower bound on the cost of adaptation for separable rules is obtained. We show that adaptability is achieved through information pooling. A tight lower bound on the amount of information pooling required for achieving rate-optimal adaptation is given. Furthermore, in a sharp contrast to the separable rules, it is shown that adaptive non-separable estimators can be superefficient at every point in the parameter spaces. The results demonstrate that information pooling is the key to increasing estimation precision as well as achieving adaptability and even superefficiency.

Keywords: Adaptability; Bayes rules; Information pooling; Minimax; Minimum risk inequalities; Nonparametric regression; Orthogonal series; Separable rules; Superefficiency; Wavelets; White noise.

AMS 1991 Subject Classification: Primary: 62G99; Secondary: 62F12, 62F35, 62M99.

*Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, tcai@wharton.upenn.edu, Phone: (215)-898-8224, Fax: (215)-898-1280. Research supported in part by NSF Grant DMS-0306576.

1 Introduction

The problem of adaptation in estimating a function globally and a function locally at a point figures prominently in the nonparametric functional estimation literature. There is an interesting distinction between the global and local estimation problems. Lepski (1990) and Brown and Low (1996b) consider the adaptation problem in estimating a function at a point. It is shown that it is impossible to achieve the minimax rate of convergence adaptively over a range of Hölder classes. That is, the estimation problem lacks adaptability. Efromovich and Low (1996a, 1996b) show that a similar phenomenon also appears in estimating certain nonlinear functionals.

On the other hand, it is well known that in the global estimation problem it is possible to achieve adaptively the minimax rate and in some cases even the minimax constant across different function classes. Indeed, one of the main goals in global estimation is to construct adaptive estimators which are simultaneously minimax over a wide range of function spaces; see, for example, Efromovich and Pinsker (1984), Cai, Low, and Zhao (2001), and Zhang (2005). It is unclear, however, why and how the adaptability is achieved. In the present paper we make the connection between adaptability and information pooling.

We begin by considering separable rules. Separable rules figure prominently in wavelet and other orthogonal series methods in contemporary nonparametric function estimation. They play a fundamental role similar to the linear estimators in more traditional function estimation literature. Separable rules are simple and intuitively appealing. More importantly, separable rules are minimax for a wide range of function classes. In deriving the minimax risk for estimating functions over Besov and Triebel classes using a wavelet basis, Donoho and Johnstone (1998) showed that the least favorable priors necessarily have independent coordinates and the Bayes minimax rules are separable. The results imply that one needs to look no further than the separable rules for the minimax estimators, provided the smoothness parameters are known.

After Section 2 in which basic notation and definitions are reviewed, adaptability of the separable rules is considered in Section 3. It is shown that if a separable rule attains the optimal rate over a Besov body, then it necessarily attains the exact same rate at every point in the Besov body. Therefore superefficiency is impossible for such an estimator at any point in the parameter space. This behavior of separable rules resembles estimators in the standard finite-dimensional normal mean problem: if an estimator is superefficient at a point, then the estimator must be penalized at a neighboring point. A direct consequence of this result is that separable rules lack adaptability; they cannot be rate-adaptive across Besov bodies. As a particular example, Bayes rules with independent priors are not rate-adaptive. Furthermore we show that separable rules must pay the minimum penalty of a logarithmic factor for adaptation and that the lower bound is sharp. Although the problem is quite different from adaptive estimation of a function at a point, the logarithmic penalty appears in both cases. The difference is that in the global estimation problem

under consideration the penalty is avoidable.

The connection between adaptability and information pooling is made in Section 4. A lower bound on the amount of information pooling required to achieve global adaptivity is derived. The results are interesting. It is shown that in order to achieve full adaptability an estimator must pool essentially at least $O(\log n)$ number of observations for estimating each individual coordinate. By using the BlockJS estimator introduced in Cai (1999) we show in Section 4.2 that the lower bound on the amount of information pooling is tight. These results together demonstrate that information pooling is the key to achieving adaptability.

Furthermore, in a sharp contrast to the separable rules, we show in Section 5 that by improving estimation accuracy through information pooling it is possible for rate-adaptive estimators to be superefficient at *every* point in the parameter spaces. This demonstrates a fundamental difference between the adaptive non-separable rules and the separable rules, as well as between the infinite-dimensional nonparametric problem and the finite-dimensional normal mean problem. It is well known that the set of superefficient points for any estimator in the finite-dimensional normal mean problem must have measure 0, but in the infinite-dimensional problem, it is possible that the set of superefficient points can be the whole parameter space. The proofs are given in Section 7.

2 Notation and definitions

In the present paper we consider the canonical infinite series version of the nonparametric function estimation problem which is exactly equivalent to the conventional white noise model. This version is also directly equivalent to nonparametric regression. See Brown and Low (1996a) and Brown, Cai, Low, and Zhang (2002). There is also a slightly less direct equivalence to nonparametric density estimation. See Nussbaum (1996) and Brown, Carter, Low and Zhang (2004).

In the conventional white noise model, we observe stochastic processes $Y_n(t)$ governed by

$$dY_n(t) = f(t)dt + n^{-1/2}dW(t), \quad 0 \le t \le 1 \tag{1}$$

where $W(t)$ is a standard Brownian motion. We wish to estimate the drift function f. The accuracy of an estimator $\hat{f}$ is measured by the mean integrated square error:

$$R(\hat{f},\ f) = E\|\hat{f} - f\|_2^2 = E\int_0^1 (\hat{f}(t) - f(t))^2\ dt. \tag{2}$$

Suppose $\{\beta_i(t),\ i \in \mathcal{I}\}$ is an orthonormal basis of $\mathcal{L}^2[0,1]$. Let $y_i = \int \beta_i(t)dY_n(t)$ and $\theta_i = \int f(t)\beta_i(t)dt$. Then the function estimation problem is exactly equivalent to the following sequence model.

Observe

$$y_i = \theta_i + n^{-1/2}z_i,\ z_i \overset{iid}{\sim} N(0,1), \quad i \in \mathcal{I} \tag{3}$$

and wish to estimate θ under the risk

$$R(\delta,\ \theta) = E_\theta \|\delta - \theta\|_{\ell^2}^2.$$

An estimator δ of the coefficient sequence θ directly provides an estimator

$$\hat{f}(t) = \sum_{i \in \mathcal{I}} \delta_i\ \beta_i(t)$$

of the function f with an isometry of risk $R(\hat{f},\ f) = R(\delta,\ \theta)$.

In the present paper our discussions will primarily focus on the Besov Spaces in the wavelet bases, although all of the results apply to, for example, the Sobolev Spaces in the Fourier basis. We will use the conventional notation in the wavelet literature and write (3) as

$$y_{j,k} = \theta_{j,k} + n^{-1/2} z_{j,k},\ \ z_{j,k} \overset{iid}{\sim} N(0,1),\ \ (j,k) \in \mathcal{J} \tag{4}$$

where the index set

$$\mathcal{J} = \{(j,\ k) :\ k = 1, \cdots, 2^j,\ \ j = 1, 2, \cdots\}.$$

The performance of a sequence of estimators $\{\delta^{(n)}\}$ is measured by its maximum risk over a parameter space $\mathcal{F}_\alpha$:

$$R_n(\delta^{(n)},\ \mathcal{F}_\alpha) = \sup_{\theta \in \mathcal{F}_\alpha} E\|\delta^{(n)} - \theta\|_{\ell^2}^2,$$

where α is some smoothness index. The benchmark is the minimax risk

$$R_n^*(\mathcal{F}_\alpha) = \inf_{\delta^{(n)}} \sup_{\theta \in \mathcal{F}_\alpha} E\|\delta^{(n)} - \theta\|_{\ell^2}^2.$$

For convenience, we will suppress the dependence of $\delta^{(n)}$ on n and omit the superscript from the notation.

In this paper, the parameter spaces of interest are the Besov Spaces which include the conventional Hölder Spaces and Sobolev Spaces as special cases. See Meyer (1992) for further details on wavelets and Besov Spaces. Define the Besov seminorm $|\cdot|_{b^\alpha_{p,q}}$ as

$$|\theta|_{b^\alpha_{p,q}} = (\sum_{j=1}^{\infty} (2^{js} (\sum_{k=1}^{2^j} |\theta_{jk}|^p)^{1/p})^q)^{1/q},$$

where $s = \alpha + 1/2 - 1/p > 0$. Then a Besov body $B^\alpha_{p,q}(M)$ is a ball under this seminorm:

$$B^\alpha_{p,q}(M) = \{\theta :\ |\theta|_{b^\alpha_{p,q}} \le M\}.$$

In the remainder of the paper, the condition $\alpha + 1/2 - 1/p > 0$ is always implicitly assumed.

It is shown in Donoho and Johnstone (1998) that the minimax rate of convergence over the Besov body $B^\alpha_{p,q}(M)$ is $n^{2\alpha/(1+2\alpha)}$. That is,

$$0 < \underline{\lim}_{n\to\infty}\ n^{2\alpha/(1+2\alpha)} R_n^*(B^\alpha_{p,q}(M)) \le \overline{\lim}_{n\to\infty}\ n^{2\alpha/(1+2\alpha)} R_n^*(B^\alpha_{p,q}(M)) < \infty.$$

3 Separable rules

Separable rules figure prominently in wavelet as well as other orthogonal series methods. They are often used as the benchmark for deriving the minimax risks or minimax estimators over Besov and other function spaces (see e.g., Donoho and Johnstone (1998) and Zhang (2005)). This fundamental role is similar to the linear estimators over more traditional function spaces in the literature.

Under the sequence model (4), an estimator $\delta = (\delta_{j,k})$ is *separable* if for all $(j,k) \in \mathcal{J}$, $\delta_{j,k}$ depends solely on $y_{j,k}$, not on any other y's. Well known examples of separable rules include term-by-term thresholding wavelet estimators and Bayes estimators derived from independent priors.

Separable rules are attractive because of their simplicity and intuitive appeal. More importantly, separable rules are minimax for a wide range of function classes. In an important paper, Donoho and Johnstone (1998) showed that the Bayes minimax rules for a Besov body $B^{\alpha}_{p,q}(M)$ are separable and furthermore the optimal separable rules are asymptotically minimax when $p \le q$ and are within a constant factor of minimax when $p > q$. Hence when smoothness parameters are known, separable rules can be optimal. Specific rate-optimal separable rules have been constructed, for example, in Delyon and Juditsky (1996) for nonparametric regression and density estimation.

3.1 Adaptive estimation

Simple separable rules can be rate-optimal over a Besov Body $B^{\alpha}_{p,q}(M)$ if the smoothness parameter α is known. A natural question is: can the optimal rate be achieved adaptively by separable rules? To answer the question, we now investigate the adaptability of separable rules. Let us begin with a simple version of the adaptation problem. Let $B^{\alpha_1}_{p_1,q_1}(M_1)$ and $B^{\alpha_2}_{p_2,q_2}(M_2)$ be two Besov bodies with $\alpha_1 \neq \alpha_2$. We call an estimator δ rate-adaptive over the two Besov bodies if δ attains the minimax rate simultaneously over them, i.e.,

$$\max_{i=1,2} \overline{\lim_{n\to\infty}}\, n^{2\alpha_i/(1+2\alpha_i)} \sup_{\theta \in B^{\alpha_i}_{p_i,q_i}(M_i)} E\|\delta - \theta\|^2_{\ell^2} < \infty.$$

Can separable rules be rate-adaptive over two Besov bodies? The answer is NO. The results below show that separable rules have their limitation; they are necessarily not rate-adaptive. We shall denote by $\mathcal{E}_t$ the class of all separable rules.

Theorem 1 *If $\delta_n \in \mathcal{E}_t$ attains the optimal rate of convergence over a Besov body $B^{\alpha}_{p,q}(M)$, then it must attain the exact same rate at every point, i.e.,*

$$0 < \underline{\lim_{n\to\infty}}\, n^{2\alpha/(1+2\alpha)} E\|\delta_n - \theta\|^2_{\ell^2} \le \overline{\lim_{n\to\infty}}\, n^{2\alpha/(1+2\alpha)} E\|\delta_n - \theta\|^2_{\ell^2} < \infty,$$

for every $\theta \in B^{\alpha}_{p,q}(M)$.

We will discuss the reasons behind Theorem 1 in Section 3.2. But first let us look at the implications of the result. A direct consequence of Theorem 1 is that separable rules are not rate-adaptive.

Corollary 1 *If $\alpha_1 \neq \alpha_2$, then*

$$\max_{i=1,2} \overline{\lim_{n\to\infty}}\, n^{2\alpha_i/(1+2\alpha_i)} \inf_{\delta\in\mathcal{E}_t} \sup_{\theta\in B^{\alpha_i}_{p_i,q_i}(M_i)} E\|\delta-\theta\|^2_{\ell^2} = \infty. \tag{5}$$

In other words, separable rules lack adaptability.

The proof of (5) is straightforward. Suppose $\alpha_1 > \alpha_2$ and a separable rule δ attains the minimax rate n^{r_2} with $r_2 = 2\alpha_2/(1+2\alpha_2)$ over the Besov body $B^{\alpha_2}_{p_2,q_2}(M_2)$. Then it follows from Theorem 1 that δ converges at the rate n^{r_2} at every point $\theta \in B^{\alpha_2}_{p_2,q_2}(M_2)$. In particular, δ converges at the rate n^{r_2} at those points in the intersection of the two Besov bodies, $B^{\alpha_1}_{p_1,q_1}(M_1) \cap B^{\alpha_2}_{p_2,q_2}(M_2)$. $B^{\alpha_1}_{p_1,q_1}(M_1) \cap B^{\alpha_2}_{p_2,q_2}(M_2)$ is always nonempty since 0 is always in the intersection. Therefore the uniform rate of convergence of δ over $B^{\alpha_1}_{p_1,q_1}(M_1)$ is at most n^{r_2} which is slower than the minimax rate $n^{2\alpha_1/(1+2\alpha_1)}$. So (5) is true.

Remark 1: In the context of Bayesian estimation, Zhao (2000) shows that independent priors must depend on n or be improper in order for the corresponding Bayes rules to achieve the optimal rate of convergence over a fixed Sobolev class. Our results above implies that independent priors cannot yield rate-adaptive Bayes rules. Hence, in order for Bayes procedures to be rate-adaptive, the priors must be more complex than the relatively simple independent priors.

Now we know that separable rules cannot achieve optimal rate adaptively. The next question is: what is the minimum cost of adaptation for separable rules? We derive below a sharp lower bound for the adaptive minimax rate of convergence for separable rules.

Theorem 2 *Suppose $\alpha_1 \neq \alpha_2$ (say, $\alpha_1 > \alpha_2$). If a separable rule δ achieves rate of convergence n^r with $r > 2\alpha_2/(1+2\alpha_2)$ over $B^{\alpha_1}_{p_1,q_1}(M_1)$ (in particular, if δ attains the minimax rate $n^{2\alpha_1/(1+2\alpha_1)}$ over $B^{\alpha_1}_{p_1,q_1}(M_1)$), then*

$$\underline{\lim_{n\to\infty}} \left(\frac{n}{\log n}\right)^{2\alpha_2/(1+2\alpha_2)} \sup_{\theta\in B^{\alpha_2}_{p_2,q_2}(M_2)} E\|\delta-\theta\|^2_{\ell^2} > 0. \tag{6}$$

That is, the rate of convergence over $B^{\alpha_2}_{p_2,q_2}(M_2)$ cannot be faster than $(n/\log n)^{2\alpha_2/(1+2\alpha_2)}$.

Therefore, the minimum cost of adaptation for the separable rules is at least a logarithmic factor. The universal threshold estimator, VisuShrink, introduced in Donoho and Johnstone (1994), achieves the convergence rate of $(n/\log n)^{2\alpha/(1+2\alpha)}$ adaptively across a range of Besov bodies $B^{\alpha}_{p,q}(M)$. Therefore the lower bound given in (6) for the adaptive minimax rate of the class of separable rules is sharp. VisuShrink is thus optimal among separable

rules in the sense that it attains the lower bound on the adaptive convergence rate for the class of the estimators.

Theorem 2 bears a strong similarity to the problem of adaptive estimation of a function at a point. It is well known that for the local estimation problem one has to pay a minimum cost of a logarithmic factor for adaptation. See Lepski (1990) and Brown and Low (1996b). The difference in these two cases is that the penalty is avoidable in the global estimation problem and unavoidable in the local estimation problem. The reason for the logarithmic penalty in (6) is that separable rules estimate each coordinate $\theta_{j,k}$ independently based solely on one individual observation $y_{j,k}$; they do not pool information contained in the observations (4) to make more informative and accurate decisions.

3.2 Reasons for lack of adaptability

To understand fully why separable rules lack adaptability, let us first review very briefly the standard univariate normal mean problem. In a univariate normal mean problem of estimating μ based on $X \sim N(\mu,\ n^{-1})$, the minimax rate of convergence over $\mathbb{R}$ under square error is n. An estimator $\hat{\mu}$ is superefficient at some point $\mu \in \mathbb{R}$ if $nE_\mu(\hat{\mu} - \mu)^2$ converges to zero. It is well known that in the univariate problem there exist estimators that are superefficient at any given point θ_0 but the estimators must "pay for" the superefficiency at θ_0 by being subefficient in a neighborhood of θ_0. The Hodges estimator is a well known example of such estimators. See Le Cam (1953) and Van der Vaart (1998). See also Brown and Low (1996b).

Under the sequence model (4), the minimax rate of convergence over the Besov body $B^\alpha_{p,q}(M)$ is $n^{2\alpha/(1+2\alpha)}$. We call an estimator δ *superefficient* at a fixed point $\theta \in B^\alpha_{p,q}(M)$ if $n^{2\alpha/(1+2\alpha)}E_\theta\|\delta - \theta\|^2_{\ell^2}$ converges to zero. Theorem 1 shows that any rate-optimal separable rule over $B^\alpha_{p,q}(M)$ cannot be superefficient at any point $\theta \in B^\alpha_{p,q}(M)$; it necessarily has a "flat" rate of convergence everywhere in $B^\alpha_{p,q}(M)$. This is not the case for non-separable rules. As we show in Section 5 that it is in fact possible for a non-separable rule to be superefficient at EVERY point in $B^\alpha_{p,q}(M)$. See Section 5 for further discussions on superefficiency.

To shed light to the reasons why separable rules lack adaptability, we give a heuristic proof of Theorem 1 here. The detailed proof is given in Section 7.

A heuristic proof of Theorem 1: Let $\delta = (\delta_{j,k})$ be a separable rule attaining the minimax rate over $B^\alpha_{p,q}(M)$. Then each $\delta_{j,k}$ can be regarded as an estimator in a univariate normal mean problem.

1. Suppose δ is superefficient at some point $\theta^* \in B^\alpha_{p,q}(M)$, i.e., δ converges faster than the minimax rate at θ^*.

2. Then as a univariate normal mean problem, many $\delta_{j,k}$ are superefficient at $\theta^*_{j,k}$ and thus each of these $\delta_{j,k}$ must be penalized in a "subefficient neighborhood" of $\theta^*_{j,k}$.

3. There exists some $\theta' \in B^\alpha_{p,q}(M)$ with coordinates $\theta'_{j,k}$ in those "subefficient neighborhoods" of $\theta^*_{j,k}$. Because δ is superefficient at θ^*, there are "too many" $\delta_{j,k}$ that are subefficient at $\theta'_{j,k}$.

4. As a consequence, δ as a whole is subefficient at θ' relative to the minimax risk over $B^\alpha_{p,q}(M)$. This contradicts the assumption that δ is rate-optimal uniformly over $B^\alpha_{p,q}(M)$.

The sketch of the proof shows that separable rules behave very similarly to the estimators in a finite-dimensional normal mean problem. That is, if an estimator is superefficient at a point in the parameter space, then it must be penalized in a neighborhood of the point of superefficiency. The main reason is that separable rules do not efficiently utilize the information contained in the sample and do not fully take advantage of the infinite-dimensional nature of the estimation problem. One can improve the estimation accuracy by information pooling.

For the infinite-dimensional problem under consideration, an estimator does not necessarily need to "pay for" superefficiency. This is one of the fundamental differences between infinite-dimensional and finite-dimensional problems. In Section 5 we show that if an estimator uses information contained in the sample (4) more efficiently, it is possible for the estimator not only to achieve adaptability uniformly over a range of the Besov bodies $B^\alpha_{p,q}(M)$, but also to be superefficient at *every* point in $B^\alpha_{p,q}(M)$.

4 Information pooling and adaptability

4.1 A lower bound on information pooling

Separable rules are necessarily not rate-adaptive, therefore in order for an estimator to achieve adaptability it must pool information contained in more than one coordinate to make more accurate decisions. A natural question is how much information pooling is necessary to achieve adaptability over Besov bodies? To answer this question we have the following result.

Theorem 3 *Let $\alpha > 0$ and let $\delta = (\delta_{j,k})$ be an estimator such that each $\delta_{j,k}$ depends on at most $h_n = o((\log n)^{2\alpha/(1+2\alpha)})$ observations. Let $\theta^\dagger \in B^\alpha_{p,q}(M)$. If*

$$\overline{\lim_{n\to\infty}}\, n^r R(\delta,\, \theta^\dagger) < \infty,$$

for some $r > 2\alpha/(1+2\alpha)$, then

$$\underline{\lim_{n\to\infty}}\, n^{2\alpha/(1+2\alpha)} \frac{h_n}{(\log n)^{2\alpha/(1+2\alpha)}} \sup_{\theta\in B^\alpha_{p,q}(M)} E\|\delta-\theta\|^2_{\ell^2} > 0.$$

In particular,

$$\underline{\lim_{n\to\infty}}\, n^{2\alpha/(1+2\alpha)} \sup_{\theta\in B^\alpha_{p,q}(M)} E\|\delta-\theta\|^2_{\ell^2} = \infty. \tag{7}$$

Therefore, in order to achieve adaptability over all Besov bodies $B^{\alpha}_{p,q}(M)$ for all $\alpha > 0$, the information pooling index h_n should be essentially at least of the order $\log n$.

In many cases the amount of information pooling varies from resolution level to resolution level, sometimes even from coefficient to coefficient within the same level. The results of Theorem 3 still holds if the condition on the amount of information pooling in the theorem is satisfied in an average sense.

Let $\delta = (\delta_{j,k})$ be an estimator and let

$$G_{j,k} = \{(l,\ m) \in \mathcal{J} : l \le j \text{ and } \delta_{j,k} \text{ depends on } y_{l,m}\}$$

be the set of indices of observations up to the level j used in the estimation of $\theta_{j,k}$. We define a sequence of information-pooling indices by

$$h_j = \text{Average } \{\text{Card}(G_{l,k}) :\ k = 1, \ldots, 2^l,\ l \le j\} = \frac{1}{2^{j+1}-1} \sum_{l=1}^{j} \sum_{k} \text{Card}(G_{l,k}),$$

where $\text{Card}(G_{l,k})$ denotes the cardinality of the set $G_{l,k}$. Here h_j can be viewed as a measure of the average amount of information pooling up to the level j. Let j_n be a sequence of integers satisfying $c_0 n^{1/(1+2\alpha)} \le 2^{j_n} \le c_1 n^{1/(1+2\alpha)}$ for some fixed constants $0 < c_0 < c_1$. If $h_{j_n}/(\log n)^{2\alpha/(1+2\alpha)} \to 0$, then, if $\overline{\lim}_{n\to\infty} n^r R(\delta,\ \theta^\dagger) < \infty$ for some $\theta^\dagger \in B^{\alpha}_{p,q}(M)$ and $r > 2\alpha/(1+2\alpha)$,

$$\underline{\lim}_{n\to\infty}\ n^{2\alpha/(1+2\alpha)} \sup_{\theta \in B^{\alpha}_{p,q}(M)} E\|\delta - \theta\|^2_{\ell^2} = \infty.$$

Therefore, if the average amount of information pooling up to the level j_n is essentially smaller than $(\log n)^{2\alpha/(1+2\alpha)}$, then the estimator has to pay for superefficiency at any point by being subefficient in a neighborhood and consequently cannot be adaptive over two Besov bodies $B^{\alpha}_{p,q}(M)$ and $B^{\alpha'}_{p',q'}(M')$ with $\alpha < \alpha'$.

One of the main tools used in the proof of Theorem 3 is a constrained risk inequality stated in Section 7.1.1. It is a generalization of the risk inequality introduced in Brown and Low (1996b) which gives a sharp lower bound for the squared error risk at one parameter point subject to having a small risk at another parameter point in a scalar-parameter univariate random variable setting. A further generalization and its applications are presented in Cai, Low and Zhao (2006). The inequalities are also related to the study of ϵ-minimax procedures and to superefficient estimation. See Brown and Low (1996b). See also Bickel (1983).

4.2 The lower bound is tight

Theorem 3 states that in order to achieve adaptability over Besov bodies $B^{\alpha}_{p,q}(M)$ for all $\alpha > 0$, the information pooling index h_n should be essentially at least of the order $\log n$. Is this lower bound tight?

Block thresholding has been shown to be an effective and convenient tool for information pooling to enhance the estimation accuracy. Recent results on block thresholding are

discussed, for example, in Hall, Kerkyacharian, and Picard (1998, 1999), Cai (1999, 2002), Cai and Silverman (2001), and Cai and Low (2005).

We will use the BlockJS estimator, introduced in Cai (1999), to show that the lower bound $O(\log n)$ on information pooling obtained in Theorem 3 is indeed tight. Furthermore, as a sharp contrast to the separable rules, we will also show in Section 5 that the BlockJS estimator has an interesting and somewhat surprising property: it is superefficient at every point in the parameter spaces. The adaptability and the superefficiency properties of BlockJS clear demonstrate the benefits of information pooling.

Among all the shrinkage estimators developed in the classical normal decision theory, the James-Stein estimator is perhaps the best-known. See James and Stein (1961) and Efron and Morris (1973). The BlockJS estimator, which was originally introduced for the nonparametric regression problem, is a blockwise application of a modified James-Stein rule. In the present sequence model setting the estimator can be defined as follows.

Let $J = [\log_2 n]$. Divide each resolution level $j < J$ into nonoverlapping blocks of approximate length $L = \log n$. Denote (jb) the b-th block at level j and $S^2_{(jb)} = \sum_{k\in(jb)} y^2_{j,k}$ the sum of squares for the block (jb). Let $\lambda_* = 4.50524$ be the root of the equation $\lambda - \log\lambda - 3 = 0$. The BlockJS estimator δ^* is given by

$$\delta^*_{j,k} = \begin{cases} (1 - \frac{\lambda_* L n^{-1}}{S^2_{(jb)}})_+ \, y_{j,k} & \text{for } k \in (jb), \;\; j < J \\ 0 & \text{for } j \geq J \end{cases} \tag{8}$$

It is shown in Cai (1999) that the BlockJS estimator (8) enjoys many desirable properties both numerically and asymptotically. In particular,

$$\overline{\lim_{n\to\infty}} \; n^{2\alpha/(1+2\alpha)} \sup_{\theta\in B^\alpha_{p,q}(M)} E\|\delta^* - \theta\|^2_{\ell^2} < \infty.$$

for all $\alpha > 0,\ p \geq 2,\ q > 0$ and $M > 0$. Therefore, BlockJS attains the optimal rate of convergence adaptively over a wide range of Besov bodies. The adaptability of the BlockJS estimator is achieved through information pooling. It pools information contained in blocks of size $\log n$ to make simultaneous shrinkage decisions for all coefficients within the same block. Furthermore, since each $\delta^*_{j,k}$ depends on at most $\log n$ observations for the BlockJS estimator, it shows that the lower bound on the amount of information pooling necessary for achieving adaptability is tight.

5 Superefficiency at a fixed point

It is shown in Section 3 that if a separable rule attains the minimax rate of convergence over a Besov body $B^\alpha_{p,q}(M)$, then it must attain exactly the same convergence rate at every point in the Besov body. Therefore, no superefficiency is possible for such an estimator. In contrast, we demonstrate in this section that adaptive non-separable estimators behave much more "intelligently". In the following theorem we use the BlockJS estimator as

an example to show that through information pooling it is possible to have estimators which are not only rate-adaptive uniformly over a wide range of parameter spaces but also superefficient at every point in the parameter spaces. This result together with Theorem 1 show a major difference in performance between the separable rules and adaptive non-separable rules.

Theorem 4 *Let δ^* denote the BlockJS estimator. Then at any fixed point $\theta \in B^\alpha_{p,q}(M)$ with $p \geq 2$ and $q < \infty$, the estimator δ^* is superefficient. That is*

$$\overline{\lim_{n\to\infty}} n^{2\alpha/(1+2\alpha)} E\|\delta^* - \theta\|^2_{\ell^2} = 0. \tag{9}$$

In other words, the BlockJS estimator is superefficient at every point in the parameter space $B^\alpha_{p,q}(M)$.

Remark 2: In fact, it can be shown that the BlockJS estimator δ^* is superefficient uniformly over any compact subset $\mathcal{C} \subset B^\alpha_{p,q}(M)$,

$$\overline{\lim_{n\to\infty}} \sup_{\theta\in\mathcal{C}} n^{2\alpha/(1+2\alpha)} E\|\delta^* - \theta\|^2_{\ell^2} = 0.$$

This result also demonstrates a fundamental difference between the infinite-dimensional problem and the classical finite-dimensional normal mean problem. In the standard finite-dimensional problem, it is well known that although superefficiency is possible, the set of superefficient points for any estimator must have measure 0. In the infinite-dimensional problem, however, it is possible to have estimators that are superefficient at every point in the parameter space. The phenomenon of superefficiency at a fixed parameter point in nonparametric function estimation has been discussed in Brown, Low, and Zhao (1997). Zhang (2005) considers fixed-parameter superefficiency in the context of an empirical Bayes estimator.

6 Concluding Remarks

Separable rules cannot achieve superefficiency at any parameter point without paying a penalty. They thus lack adaptability. The difficulty of separable rules is due to the relative inaccuracy with which individual coordinates are estimated when the smoothness parameter is unknown. Information pooling is shown to be the key to increase estimation precision and achieve adaptability. In order to achieve full adaptability an estimator must use at least $O(\log n)$ number of observations for estimating each individual coordinate. The lower bound on information pooling is tight. Moreover, in a sharp contrast to the separable rules, it is shown that adaptive non-separable estimators can be superefficient at every point in the parameter spaces.

Besides block thresholding, empirical Bayes is another effective way of pooling information to achieve adaptability. Johnstone and Silverman (2005) and Zhang (2005) demonstrate that adaptability can be achieved by relatively simple empirical Bayes procedures.

7 Proofs

7.1 Proof of Theorems 1 and 2

We first introduce the following constrained risk inequality.

7.1.1 A general constrained risk inequality

Let X be a (vector-valued) random variable having distribution $P_{\theta_1,\xi}$ with density $f_{\theta_1,\xi}$, or distribution $P_{\theta_2,\xi}$ with density $f_{\theta_2,\xi}$, with respect to a measure λ. Here the parameter of interest is θ and ξ is some fixed nuisance parameter. Suppose $\theta_i = (\theta_{i,1}, ..., \theta_{i,K}) \in I\!R^K$ $(i = 1, 2)$. For any estimator δ of θ based on X its risk is defined by

$$R(\delta,\ \theta) = E\|\delta(X) - \theta\|_{\ell^2}^2 = \int \sum_{k=1}^{K} |\delta_k(x) - \theta_k|^2 f_\theta(x)\lambda(dx)$$

Denote by $r(x) = f_{\theta_2,\xi}(x)/f_{\theta_1,\xi}(x)$ the ratio of the two density functions. ($r(x) = \infty$ for some x is possible, with the obvious interpretation $r(x)f_{\theta_1,\xi}(x) = f_{\theta_2,\xi}(x)$.) Denote

$$D = \|\theta_2 - \theta_1\|_{\ell^2} = (\sum_{k=1}^{K} |\theta_{2,k} - \theta_{1,k}|^2)^{1/2} \tag{10}$$

and

$$\Delta = \Delta(\theta_1,\ \theta_2) = (E_{\theta_1}(r^2(X)))^{1/2}. \tag{11}$$

The following result gives a lower bound for $R(\delta,\ \theta_2)$ under the constraint of $R(\delta,\ \theta_1) \le \epsilon^2$.

Theorem 5 *Suppose $R(\delta,\ \theta_1) \le \epsilon^2$ and $D > \epsilon\Delta$, then*

$$R(\delta,\ \theta_2) \ge (D - \epsilon I)^2 \ge D^2(1 - \frac{2\epsilon\Delta}{D}). \tag{12}$$

Remark 3: The constrained risk inequality (12) is a generalization of the risk inequality introduced in Brown and Low (1996b). A further generalization with proof is presented in Cai, Low and Zhao (2006).

7.1.2 A preparatory result

Theorems 1 and 2 are consequences of the following result.

Proposition 1 *Suppose $A_n \to \infty,\ n/\log A_n \to \infty$. Let $\theta^\dagger \in B^\alpha_{p,q}(M)$ be fixed and let $\delta = (\delta_{j,k}) \in \mathcal{E}_t$ be a separable rule. If*

$$\overline{\lim_{n\to\infty}}\, n^{2\alpha/(1+2\alpha)} A_n R(\delta,\ \theta^\dagger) < \infty,$$

then

$$\underline{\lim_{n\to\infty}} \left(\frac{n}{\log A_n}\right)^{2\alpha/(1+2\alpha)} \sup_{\theta\in B^\alpha_{p,q}(M)} R(\delta,\ \theta) > 0.$$

Proof of Proposition 1: Without loss of generality, we assume $\theta^\dagger = 0$. Denote $w = 2\alpha/(1+2\alpha)$. Since $\overline{\lim}_{n\to\infty} n^w A_n R(\delta,\ 0) < \infty$, there exists constants $c, N > 0$ such that for all $n \geq N$,

$$R(\delta,\ 0) \equiv \sum_{(j,k)\in\mathcal{J}} E_0(\delta_{j,k}(y_{j,k}))^2 \leq cn^{-w}A_n^{-1}. \tag{13}$$

Let j_0 be the smallest integer satisfying $2^{j_0} \geq 2(n/\log A_n)^{1-w}$. Let

$$G_n = \{(j,k) \in \mathcal{J} :\ j \leq j_0,\ \text{and } E_0(\delta_{j,k}(y_{j,k}))^2 \leq cn^{-1}A_n^{-1/2}\} \tag{14}$$

It is easy to see from (13) that the number of indices $(j,k) \in \mathcal{J}$ with $j \leq j_0$ that are not in G_n is at most $n^{1-w}A_n^{-1/2}$. Therefore the cardinality K_n of the set G_n is at least $(n/\log A_n)^{1-w}$ when n is sufficiently large. Let

$$\theta_{j,k} = \begin{cases} c_1(\log A_n/n)^{1/2} & \text{if } (j,k) \in G_n \\ 0 & \text{if } (j,k) \notin G_n \end{cases} \tag{15}$$

where $c_1 > 0$ is some constant. We now need the following lemma. The proof is straightforward and is thus omitted.

Lemma 1 *There exists some constant $c_2 > 0$ such that for all $0 \leq c_1 \leq c_2$, the sequence $\theta = (\theta_{j,k})$ defined in (15) belongs to the Besov body $B^\alpha_{p,q}(M)$.*

Now return to the proof of Proposition 1. Denote the density function of $N(\theta,\sigma^2)$distribution by $\phi(x;\theta,\sigma^2)$. Using the notation in Section 7.1.1, for $P_{\theta_i,\sigma^2} = N(\theta_i,\sigma^2)$ simple calculus shows that

$$\Delta(\theta_1,\theta_2) = \left(E_{\theta_1}\frac{\phi^2(X;\theta_2,\sigma^2)}{\phi^2(X;\theta_1,\sigma^2)}\right)^{\frac{1}{2}} = \exp\{\frac{(\theta_2-\theta_1)^2}{2\sigma^2}\}.$$

Let $c_2 > 0$ be given as in Lemma 1. Fix $0 < c_1 < \min(1/\sqrt{2},\ c_2)$. Then $\theta = (\theta_{j,k})$ defined in (15) is in $B^\alpha_{p,q}(M)$. For each $(j,k) \in G_n$, we then have

$$\Delta(0,\ \theta_{j,k}) = \exp\{\frac{n\theta_{j,k}^2}{2}\} = A_n^{c_1^2/2} \leq A_n^{1/4}$$

and $\epsilon = c^{1/2}n^{-1/2}A_n^{-1/4}$. Now (12) yields

$$\begin{aligned} E_{\theta_{j,k}}(\delta_{j,k}(y_{j,k}) - \theta_{j,k})^2 &\geq \theta_{j,k}^2(1 - \frac{2\epsilon\Delta}{|\theta_{j,k}|}) \geq c_1\frac{\log A_n}{n}\cdot(1 - \frac{2c^{1/2}n^{-1/2}A_n^{-1/4}A_n^{1/4}}{c_1n^{-1/2}(\log A_n)^{1/2}}) \\ &= c_1\frac{\log A_n}{n}\cdot(1 - \frac{2c^{1/2}}{c_1(\log A_n)^{1/2}}) \end{aligned}$$

Therefore, for $\theta = (\theta_{j,k})$ defined in (15),

$$R(\delta,\ \theta) \geq \sum_{(j,k)\in G_n} E_{\theta_{j,k}}(\delta_{j,k}(y_{j,k}) - \theta_{j,k})^2 \geq K_n\frac{\log A_n}{n}\ (c_1 + o(1)) \tag{16}$$

$$\geq (\frac{\log A_n}{n})^{2\alpha/(1+2\alpha)}(c_1 + o(1)). \tag{17}$$

Hence,

$$\underline{\lim}_{n\to\infty}\left(\frac{n}{\log A_n}\right)^{2\alpha/(1+2\alpha)}\sup_{\theta\in B^{\alpha}_{p,q}(M)} R(\delta,\ \theta)\geq c_1>0. \quad \blacksquare$$

Remark 4: The proof shows that if a separable rule δ is superefficient at some $\theta^{\dagger}\in B^{\alpha}_{p,q}(M)$, then, coordinatewise as an estimate of $\theta^{\dagger}_{j,k}$, $\delta_{j,k}$ must be superefficient at a large number of coordinates of $\theta^{\dagger}$. This in turn forces the estimators $\delta_{j,k}$ to be subefficient in a small neighborhood of each of those coordinates. As a direct consequence the sum of mean squared errors of δ is large in a neighborhood of $\theta^{\dagger}$ which makes the rate of convergence of δ over $B^{\alpha}_{p,q}(M)$ suboptimal. As we see from Theorem 4 that this phenomenon can be avoided by using non-separable rules.

7.1.3 Proof of Theorems 1 and 2

With the preparations given in Sections 7.1.1 and 7.1.2, Theorems 1 and 2 are now easy to prove.

Proof of Theorem 1: Suppose that δ attains the minimax rate over the Besov body $B^{\alpha}_{p,q}(M)$ and δ is superefficient at some point $\theta^{\dagger}\in B^{\alpha}_{p,q}(M)$, i.e., $R(\delta,\ \theta^{\dagger})$ converges to 0 faster than the rate $n^{2\alpha/(1+2\alpha)}$. Then there exists $A_n\to\infty$ and $n/\log A_n\to\infty$ such that $\overline{\lim}_{n\to\infty} n^{2\alpha/(1+2\alpha)}A_nR(\delta,\ \theta^{\dagger})<\infty$. Now Proposition 1 yields that

$$\underline{\lim}_{n\to\infty}\left(\frac{n}{\log A_n}\right)^{2\alpha/(1+2\alpha)}\sup_{\theta\in B^{\alpha}_{p,q}(M)} R(\delta,\ \theta)>0.$$

Hence, $\underline{\lim}_{n\to\infty} n^{2\alpha/(1+2\alpha)}\sup_{\theta\in B^{\alpha}_{p,q}(M)} R(\delta,\ \theta)=\infty$, which contradicts the assumption that δ is rate optimal over $B^{\alpha}_{p,q}(M)$. $\blacksquare$

Proof of Theorem 2: First note that $0\in B^{\alpha_1}_{p_1,q_1}(M_1)\cap B^{\alpha_2}_{p_2,q_2}(M_2)$. Since δ attains the rate n^r over $B^{\alpha_1}_{p_1,q_1}(M_1)$, δ converges at least at the rate n^r at 0. Let $A_n=n^{r-2\alpha_2/(1+2\alpha_2)}$. Then

$$\overline{\lim_{n\to\infty}}\ n^{2\alpha_2/(1+2\alpha_2)}A_nR(\delta,\ 0)<\infty.$$

Now Theorem 2 follows from Proposition 1 with $A_n=n^{r-2\alpha_2/(1+2\alpha_2)}$. $\blacksquare$

7.2 Proof of Theorem 3

Again we assume, without loss of generality, the point of superefficiency $\theta^{\dagger}=0$. Since $\overline{\lim}\, n^rR(\delta,\ 0)<\infty$, for sufficiently large n,

$$R(\delta,\ 0)\equiv\sum_{(j,k)\in\mathcal{J}} E_0\delta^2_{j,k}\leq cn^{-r}. \tag{18}$$

where $c>0$ is some fixed constant. Let j_0 be the smallest integer satisfying $2^{j_0}\geq 6(n/\log n)^{1/(1+2\alpha)}$. Let $A_n=n^{r-2\alpha/(1+2\alpha)}$ and let

$$G_1=\{(j,k)\in\mathcal{J}:\ j\leq j_0,\ \text{and}\ E_0\delta^2_{j,k}\leq cn^{-1}A_n^{-1/2}\}. \tag{19}$$

Then it is easy to see that for sufficiently large n the cardinality of G_1 is at least 2^{j_0}.

We call two indices $(j,k),\ (l,m) \in \mathcal{J}$ *related* if $\delta_{j,k}$ depends on $y_{l,m}$ or $\delta_{l,m}$ depend on $y_{j,k}$, and *unrelated* otherwise. Define

$$G_2 = \{(j,k) \in G_1 : \text{the number of } (l,m) \in G_1 \text{ with } \delta_{l,m} \text{ depending on } y_{j,k} \text{ is at most } 2h_n\}.$$

Then it is easy to see that $\mathrm{Card}(G_2) \geq 2^{j_0-1}$. Let $(j,k) \in G_2$. Define $I_{j,k}$ be the set of all indices in G_2 that are related to (j,k). Then $\mathrm{Card}(I_{j,k}) \leq 3h_n$ for all $(j,k) \in G_2$. This implies that there exists a subset $G_n' \subseteq G_2$ such that all indices in G_n' are mutually unrelated and

$$K_n' \equiv \mathrm{Card}(G_n') \geq \mathrm{Card}(G_2)/(3h_n) \geq h_n^{-1}(n/\log n)^{1/(1+2\alpha)}.$$

Define θ same as in (15) with $A_n = n^{r-2\alpha/(1+2\alpha)}$ and G_n replaced by G_n'. Then again for a sufficiently small constant $0 < c_1 < 1/4$, θ is in $B_{p,q}^{\alpha}(M)$. Fix $(j,\ k) \in G_n'$. Let $Y^{(j,k)} = (y_{l,m} : \delta_{j,k} \text{ depends on } y_{l,m})$ be the vector of observations used in estimating $\theta_{j,k}$. Without loss of generality, let us put $y_{j,k}$ as the first element of the vector $Y^{(j,k)}$. Then under θ the mean of $Y^{(j,k)}$ can be written as $(\theta_{j,k},\ 0, \ldots, 0)$, since all indices in G_n' are mutually unrelated and all coordinates in θ with indices not in G_n' are 0.

Now applying the constrained risk inequality (12) with $\theta_1 = 0$ and $\theta_2 = \theta_{j,k}$, we have, after some algebra,

$$E_\theta(\delta_{j,k}(Y^{(j,k)}) - \theta_{j,k})^2 \leq c_1 \frac{\log A_n}{n} \cdot (1 - \frac{2c^{1/2}}{c_1(\log A_n)^{1/2}}).$$

Since $K_n' \geq h_n^{-1}(n/\log n)^{1/(1+2\alpha)}$, so

$$\begin{aligned} R(\delta,\ \theta) &\geq \sum_{(j,k)\in G_n'} E_{\theta_{j,k}}(\delta_{j,k}(y_{j,k}) - \theta_{j,k})^2 \geq K_n' \frac{\log A_n}{n}\ (c_1 + o(1)) \\ &\geq h_n^{-1}(\frac{\log n}{n})^{2\alpha/(1+2\alpha)}(c_2 + o(1)). \end{aligned}$$

where the constant $c_2 = c_1(r - 2\alpha/(1+2\alpha)) > 0$. Hence,

$$\underline{\lim}_{n\to\infty}\ h_n \left(\frac{n}{\log n}\right)^{2\alpha/(1+2\alpha)} R(\delta,\ \theta) \geq c_2 > 0.$$

The proof of (7) is similar. ∎

7.3 Proof of Theorem 4

Let $\theta \in B_{p,q}^{\alpha}(M)$ be fixed. Denote by $\theta_{j\cdot} = (\theta_{j,k})_{k=1}^{2^j}$ the coefficient vector at level j. Let $\gamma_j = 2^{j(\alpha+1/2-1/p)}\|\theta_{j\cdot}\|_p$. Since $\theta \in B_{p,q}^{\alpha}(M)$,

$$|\theta|_{b_{p,q}^s} = (\sum_{j=1}^{\infty} \gamma_j^q)^{1/q} \leq M.$$

So, $\gamma_j \to 0$ as $j \to \infty$. Let $\rho_j = \sup_{j' \geq j} \gamma_{j'}$. Then ρ_j also tends to 0 as $j \to \infty$. Let J_n be the largest integer satisfying $2^{J_n} \leq \rho_{J_n}^{2/(1+2\alpha)} n^{1/(1+2\alpha)}$. It follows from the BP Oracle Inequality in Cai (1999), the risk of the BlockJS estimator $\hat{\theta}^*$ at each resolution level $j < J$ can be bounded as

$$\sum_k E(\hat{\theta}^*_{j,k} - \theta_{j,k})^2 \leq \sum_b (\beta^2_{(jb)} \wedge \lambda_* L n^{-1}) + 2^{j+1} n^{-2} \tag{20}$$

where $\beta^2_{(jb)} = \sum_{k \in (jb)} \theta^2_{j,k}$ is the sum of the squared coefficients within the block (jb). Now (20) yields

$$\begin{aligned} E\|\hat{\theta}^* - \theta\|^2_{\ell^2} &\leq \sum_{j=1}^{J-1} \sum_b (\beta^2_{(jb)} \wedge \lambda_* L n^{-1}) + 2n^{-1} + \sum_{j=J}^{\infty} \sum_{k=1}^{2^j} \theta^2_{jk} \\ &\leq \sum_{j=J_n}^{J-1} \sum_b \beta^2_{(jb)} + \sum_{j=1}^{J_n - 1} \sum_b \lambda_* L n^{-1} + 2n^{-1} + C n^{-2\alpha} \\ &\leq \sum_{j=J_n}^{J-1} \sum_{k=1}^{2^j} \theta^2_{j,k} + C \rho_{J_n}^{2/(1+2\alpha)} n^{-2\alpha/(1+2\alpha)} + 2n^{-1} + C n^{-2\alpha}. \end{aligned} \tag{21}$$

Note that the following elementary inequalities on two different norms hold:

$$\|x\|_{p_2} \leq \|x\|_{p_1} \leq m^{\frac{1}{p_1} - \frac{1}{p_2}} \|x\|_{p_2}, \quad \text{for } x \in I\!R^m \text{ and } 0 < p_1 \leq p_2 \leq \infty. \tag{22}$$

It follows from (22) that

$$\|\theta_{j\cdot}\|_{\ell^2} \leq 2^{j(1/2 - 1/p)} \|\theta_{j\cdot}\|_{\ell^p} = 2^{-\alpha j} \gamma_j.$$

So,

$$\begin{aligned} \sum_{j=J_n}^{J-1} \sum_{k=1}^{2^j} \theta^2_{j,k} &\leq \sum_{j=J_n}^{J-1} \sum_{k=1}^{2^j} 2^{-2\alpha j} \gamma_j^2 \leq \rho^2_{J_n} \cdot C \rho_{J_n}^{-4\alpha/(1+2\alpha)} n^{-2\alpha/(1+2\alpha)} \\ &= C \rho_{J_n}^{2/(1+2\alpha)} n^{-2\alpha/(1+2\alpha)}. \end{aligned} \tag{23}$$

Since $\rho_{J_n} \to 0$ as $n \to \infty$, it follows from (21) and (23) that

$$\overline{\lim_{n \to \infty}} \, n^{2\alpha/(1+2\alpha)} E\|\hat{\theta}^* - \theta\|^2_{\ell^2} = 0. \quad \blacksquare$$

Acknowledgment

It is a pleasure to acknowledge helpful comments by Larry Brown and Iain Johnstone which improved the presentation of the paper.

References

[1] Bickel, P. J. (1983). Minimax estimation of the mean of a normal distribution subject to doing well at a point. In Rizvi, Rustagi and Siegmund (eds), *Recent Advances in Statistics*. Academic Press, 511-528.

[2] Brown, L.D., Cai, T., Low, M.G. and Zhang, C. (2002). On asymptotic equivalence of white noise model and nonparametric regression with random designs. *Ann. Statist.* **30**, 688-707.

[3] Brown, L.D. and Low, M.G. (1996a). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24**, 2384-2398.

[4] Brown, L.D. and Low, M.G. (1996b). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.* **24**, 2524-2535.

[5] Brown, L.D., Carter, A.V., Low, M.G. and Zhang, C. (2004). Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift. *Ann. Statist.* **32**, 2074-2097.

[6] Brown, L.D., Low, M.G., and Zhao, L. (1997). Superefficiency in nonparametric functional estimation. *Ann. Statist.* **25**, 2607-25.

[7] Cai, T. (1999). Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *Ann. Statist.* **27**, 898-924.

[8] Cai, T. (2002). On adaptive estimation of a derivative and other related linear inverse problems. *J. Stat. Planning and Inf.* **108**, 329-349.

[9] Cai, T. and Low, M. (2005). Nonparametric function estimation over shrinking neighborhoods: Superefficiency and adaptation. *Ann. Statist.* **33**, 184-213.

[10] Cai, T., Low, M. and Zhao, L. (2001). Sharp adaptive estimation by a blockwise method. Unpublished manuscipt.

[11] Cai, T., Low, M.G., and Zhao, L. (2006). Tradeoffs between global and local risks in nonparametric function estimation. *Bernoulli*, to appear.

[12] Cai, T. and Silverman, B.W. (2001). Incorporating information on neighboring coefficients into wavelet estimation. *Sankhya* (ser. B) **63**, 127-148.

[13] Delyon, B. and Juditsky, A. (1996). On minimax wavelet estimators. *Appl. Comput. Harm. Anal.* **3**, 215-228.

[14] Donoho, D.L. and Johnstone, I.M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **81**, 425-455.

[15] Donoho, D.L. and Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90**, 1200-1224.

[16] Donoho, D.L. and Johnstone, I.M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26**, 879-921.

[17] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B* **57**, 301-369.

[18] Donoho, D. L. and Low, M. G. (1992), Renormalization exponents and optimal pointwise rates of convergence. *Ann. Statist.* **20**, 944-970.

[19] Efromovich, S. and Low, M.G. (1996a). On Bickel and Ritov's conjecture about adaptive estimation of the integral of the square of density derivative. *Ann. Statist.* **24**, 682-686.

[20] Efromovich, S. and Low, M.G. (1996b). On optimal adaptive estimation of a quadratic functional. *Ann. Statist.* **24**, 1106-1125.

[21] Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors – an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68**, 117 –130.

[22] Hall, P., Kerkyacharian, G. and Picard, D. (1998). Block threshold rules for curve estimation using kernel and wavelet methods. *Ann. Statist.* **26**, 922-942.

[23] Hall, P., Kerkyacharian, G. and Picard, D. (1999). On the minimax optimality of block thresholded wavelet estimators. *Statist. Sinica* **9**, 33-50.

[24] James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 361-380. Univ. California Press.

[25] Johnstone, I. M. and Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Statist.* **33**, 1700-1752.

[26] Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes estimate. *Univ. Calif. Publ. Statist.* **1**, 277-330.

[27] Lepski, O. V. (1990), On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35**, 454-466.

[28] Nussbaum, M. (1996). Asymptotic equivalence of density estimation and white noise. *Ann. Statist.* **24**, 2399-2430.

[29] Van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press.

[30] Zhang, C. (2005). General empirical Bayes wavelet methods and exactly adaptive minimax estimation. *Ann. Statist.* **33**, 54-100.

[31] Zhao, L. (2000). Bayesian aspects of some nonparametric problems. *Ann. Statist.* **28**, 532-552.